Current Directions in Psychological Science
2022, Vol. 31(1) 76–87

DOI: 10.1177/09637214211056906
www.psychologicalscience.org/CDPS

Psychological Measurement in the Information Age: Machine-Learned Computational Models

Sidney K. D'Mello[1,2], Louis Tay[3], and Rosy Southwell[1]
[1]Institute of Cognitive Science, University of Colorado Boulder; [2]Department of Computer Science, University of Colorado Boulder; and [3]Department of Psychological Sciences, Purdue University

Abstract
Psychological science can benefit from and contribute to emerging approaches from the computing and information sciences driven by the availability of real-world data and advances in sensing and computing. We focus on one such approach, machine-learned computational models (MLCMs)—computer programs learned from data, typically with human supervision. We introduce MLCMs and discuss how they contrast with traditional computational models and assessment in the psychological sciences. Examples of MLCMs from cognitive and affective science, neuroscience, education, organizational psychology, and personality and social psychology are provided. We consider the accuracy and generalizability of MLCM-based measures, cautioning researchers to consider the underlying context and intended use when interpreting their performance. We conclude that in addition to known data privacy and security concerns, the use of MLCMs entails a reconceptualization of fairness, bias, interpretability, and responsible use.

Keywords
artificial intelligence, big data, computer science, psychological assessment

If measurement is the cornerstone of science, psychological science has accomplished a lot. Psychological scientists have designed clever experiments to measure complex social phenomena, honed the measurement of ill-defined constructs to a precise science, made inferences about the mind by probing behavior, begun to delve into the brain, and applied findings to improve the human condition. Meanwhile, the trifecta of the Information Age—new, improved, and cost-effective sensing; anywhere, anytime computing; and a new generation of people who have grown up in a digital world—has led to a data and computing revolution that has enhanced multiple research areas and created new ones (e.g., computational social science, cyber-physical systems, quantitative biology). Can such advances similarly enhance psychological science? We think so and describe how the core of psychological science—psychological measurement—can benefit from an Information Age update.

Consider one simplified view of psychological measurement: measurement = data + inference. The data typically come from humans (e.g., posts on social media) and are converted to a structured format (e.g., human coders count the number of pronouns). Computers can automate and scale up this task and discover complex associations in the data, revealing multivariate interactions and nonlinearities. However, they cannot make meaning of any patterns they discover, at least not in any deep sense. Researchers rely on human knowledge and expertise to make inferences from data. Even when measurement is automated, for example, as in computerized adaptive testing (Wainer et al., 2000), the items and inference are preprogrammed into the computer.

But what if researchers could design computers to learn how to make human-like inferences from data? The resultant measure would combine the pattern-finding

Corresponding Author:
Sidney K. D'Mello, Institute of Cognitive Science and Department of Computer Science, University of Colorado Boulder
Email: sidney.dmello@colorado.edu

prowess of computers with the inferencing abilities of humans—and would have transformative impacts. Such a measure would enable the analysis of relatively unstructured data sets (e.g., images or text on the Internet) with the scope and scale to address thorny issues of reproducibility and generalizability. By leveraging modern sensing and analysis capabilities, these measures could focus on real-world human behavior rather than curated responses. Measurement could also be done in real time, which would open the door for just-in-time interventions, individualized experimental manipulations, and discoveries currently precluded by measurement latencies. The measures would potentially be more objective provided that bias is mitigated in their design. Because the measures would be learned, not preprogrammed, analysis of the measures themselves could deepen understanding of the underlying phenomena.

If this all seems too fanciful, rest assured that there is a systematic approach to developing such measures. It is called *computational modeling*, representation of a phenomenon in silico (i.e., using computer software or simulation). This is not an advance in itself; the novelty is that the computational models are directly learned (i.e., constructed) from data rather than preprogrammed, as we elaborate next.

Machine-Learned Computational Models

A computational model is a computer program that produces a desired output given input. Applied to psychological measurement, this entails converting input data into higher-level representations, or *features*, usable by a computer, and then transforming these features into measurement estimates (i.e., output) via various algorithmic *structures*. For example, a computational model of mind wandering during reading (Faber et al., 2018) based on eye tracking can map features, such as the number and duration of gaze fixations, onto estimates of mind wandering using one of the structures in Figure 1a.

Computational models differ in how features, structure, and parameters (e.g., regression weights) are specified. Traditionally, human experts preprogrammed the models by specifying all of these components (Fig. 2), as in the classic GOMS (goals, operators, methods, and selection) models in human-factors research (Card et al., 1983). Such *handcrafted* models are rare because of difficulties in specifying a generalizable set of parameters (among other factors). An intermediate approach used in developing traditional psychological models (e.g., item-response-theory models used in assessment and classic Bayesian models of cognition) is to prespecify the features and structure but have computer algorithms learn the parameters from data.

But what about complex, poorly understood phenomena, for which neither the model structure nor the parameters can be prespecified? Using *supervised machine learning*, it is possible for the computer program to learn both from data (Jordan & Mitchell, 2015). Starting with a set of *training examples*, which link features with corresponding *annotations* (e.g., human ratings), the program constructs a model by identifying patterns in the training data. After training is complete, the resultant *machine-learned computational model* (MLCM; Fig. 2) produces computer estimates (i.e., measurements) for new input data (without annotations).

In the case of our mind-wandering example, training data are collected by tracking eye gaze (to compute features) and self-reports of mind wandering (annotations) as participants read. Training examples are created by aligning the gaze features with the mind-wandering reports over a temporal window (e.g., a page), and then supervised-learning methods are applied to generate an MLCM, which produces estimates of mind wandering based on gaze features.

What are these supervised-learning techniques? Linear regression is one potential example of a minimalist technique. However, in the psychological sciences, where the goal is *explanation*, the models are typically fit on the entire data set, and the emphasis is on statistical significance of the coefficients (Yarkoni & Westfall, 2017). For machine learning, the goal is instead *prediction*, and the focus is on the extent to which MLCM outputs align with some measure of "ground truth" when applied to *holdout* data (i.e., data different from training data), including data from different people, paradigms, populations, and contexts. In other words, in the case of machine learning, the focus is on whether the model is *generalizable* (e.g., whether it accurately predicts self-reports of mind wandering among a different set of people reading a new text).

A highly accurate model might overfit the training data and perform poorly on holdout data (low generalizability), whereas a highly generalizable model might underfit the data (low accuracy). Because regression and its variants (e.g., generalized linear models) are limited in both respects, researchers have developed numerous approaches to improving accuracy (e.g., modeling nonlinearity and feature interactivity) and generalizability (e.g., using an ensemble of models and penalizing those with more parameters). As Figure 1 indicates, the resultant models have different representations (e.g., probabilities, parameter weights), structures (e.g., equations, rules, networks of artificial neurons), and assumptions (e.g., some assume feature independence, whereas feature interdependence is critical in others). But they are all computer programs.

a

Regression Structure

$$MW = B_0 + B_1 \times NFix + B_2 \times FixDur$$

Rule-Based (Tree-Based) Structure

if [NFix < $T1$ and FixDur > $T2$]; then [MW = true]

(Naive) Bayesian Structure (Simplified)

$$P(MW \mid NFix = x, FixDur = y) = [P(NFix = x \mid MW) \times P(FixDur = y \mid MW) \times P(MW)]/P(NFix = x, FixDur = y)$$

Neural Network Representation (Weight Matrix)

$$\begin{bmatrix} w_{NFix\_h1} & w_{FixDur\_h1} \\ w_{NFix\_h2} & w_{FixDur\_h2} \end{bmatrix} \begin{matrix} w_{h1\_MW} \\ w_{h2\_MW} \end{matrix}$$

b

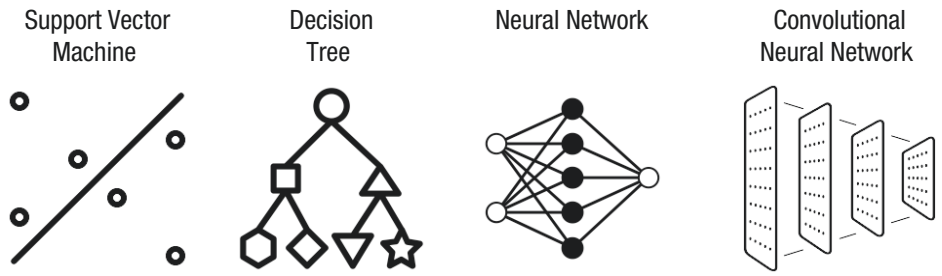

Fig. 1. Examples of different structures and representations for machine-learned computational models (MLCMs): (a) example structures for MLCMs of mind wandering (MW) based on two eye-gaze features, number of fixations (NFix) and fixation duration (FixDur), and (b) graphical representations of some common types of MLCMs illustrating differences in how the various approaches encode the data, for example, as a decision boundary (support vector machine), in a flowchart-like structure (decision tree), and in node-link assemblies (neural networks). B = parameter; T = threshold; P = probability, W = weight; h1 and h2 = hidden nodes 1 and 2.

These standard machine-learning approaches can be contrasted with *deep* (*neural*) *learning* (Jordan & Mitchell, 2015), which combines massive data (e.g., the entirety of English Wikipedia), rather than (or in addition to) prespecified features, with the requisite computing (e.g., thousands of parallel processors) and advanced algorithms to process the data, which results in an increase in MLCM complexity (up to billions of parameters) and performance improvements. One innovation is *representational learning*, in which the features themselves are learned from raw data rather than being prespecified. An extension is *end-to-end learning*, in which everything (features, structure, and parameters) is learned simultaneously from raw data. For example, rather than using human-engineered features, a model of mind wandering during reading might automatically extract internal representations most useful for predicting mind wandering from raw gaze data. Another extension is *fine tuning*, pretraining a model on massive data in a domain-agnostic fashion (e.g., with large volumes of gaze data from multiple studies without any annotations of mind wandering) to extract internal representations and then adapting the model for a given domain using a small amount of annotated data.

As Figure 2 indicates, computational models can be broadly divided into *explanatory* models, for which the primary aim is understanding the underlying mechanisms, and *predictive* models, for which accurate and generalizable predictions are the main goal. MLCMs fall into the predictive family in that they have fewer theoretical commitments than explanatory models and are more bottom-up and data driven. As a result, MLCMs with very different structures can yield similar predictions, which limits their ability to provide causal or mechanistic explanations. However, because they are powerful, fine-grained predictive machines, MLCMs can be useful tools for scientific inquiry (in addition to their

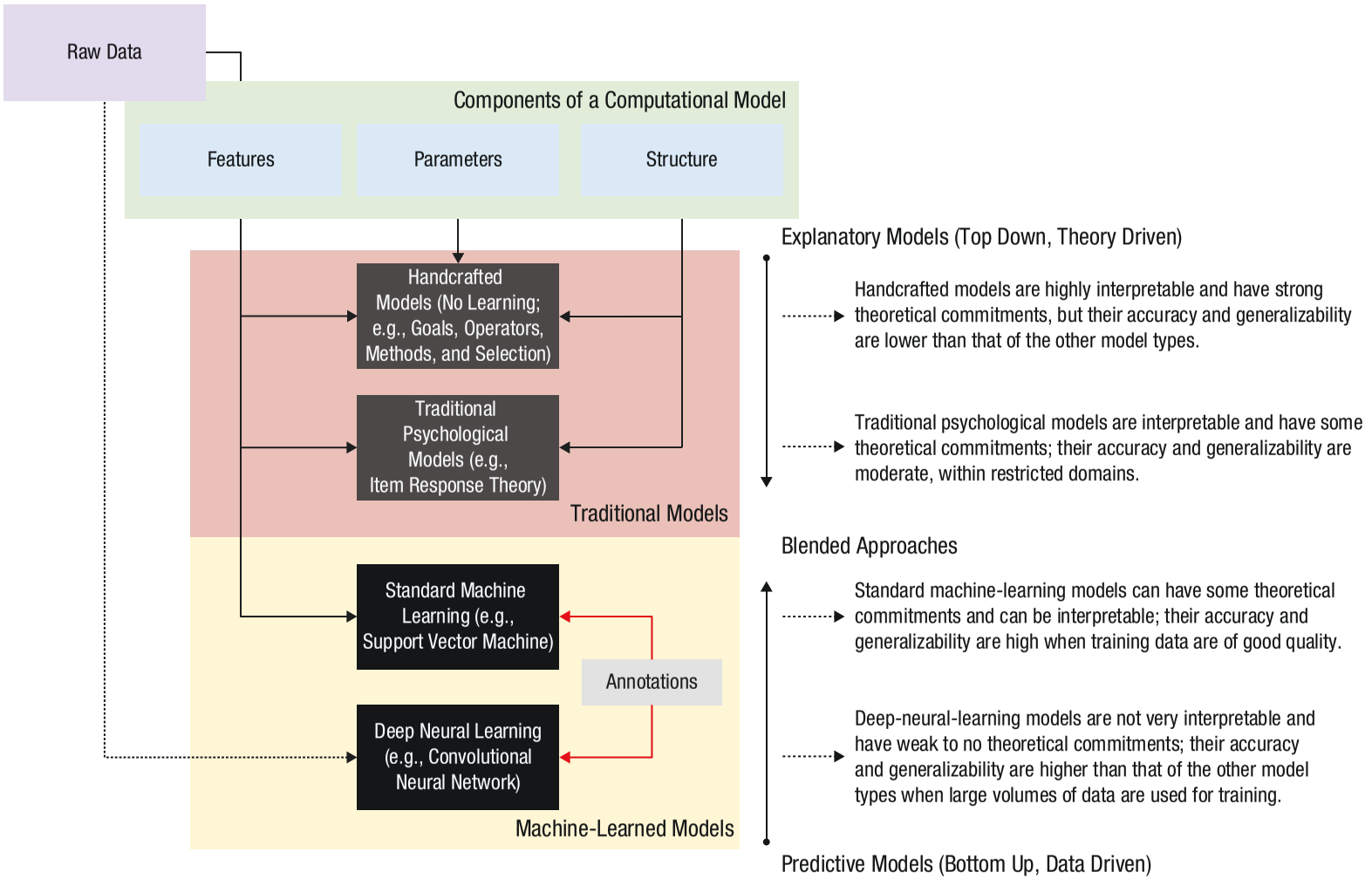

Fig. 2. The four main approaches to computational modeling. The approaches differ in whether features, parameters, and structure are prespecified. Handcrafted and traditional psychological models are more explanatory, whereas standard machine-learning and deep-neural-learning models are more predictive. Note that the approaches are not mutually exclusive and can be combined in multiple ways; explanation and prediction goals are combined in blended approaches. In the case of deep-neural-learning models, raw data can be input directly; for the other model types, features are first prespecified and then computed from raw data and used as inputs for learning. Parameters are prespecified for handcrafted models only, and model structure is prespecified for both types of traditional models. Annotations are labels provided by humans to guide the machine-learning process by providing a supervisory signal. They are needed in the model-training phase for supervised machine learning. Technically, a response variable (not shown) is needed for traditional models, but such variables are not considered to be annotations. Dotted, solid black, and solid red lines indicate requirements for minimal, typical, and substantial human knowledge engineering, respectively.

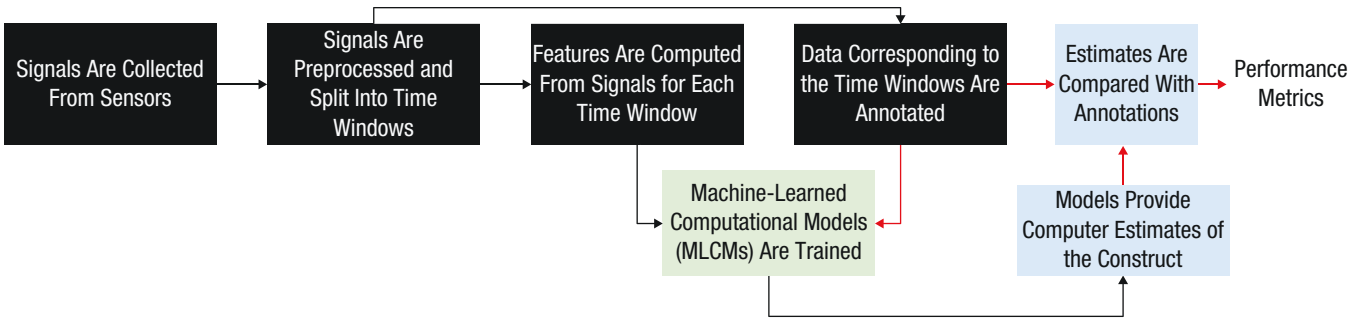

Fig. 3. The basic pipeline for training standard machine-learned computational models (MLCMs). The arrows denote the flow of information processing; red arrows denote steps that are involved in the training process only and are skipped once MLCMs have been trained.

use in assessment and intervention). For example, they can be designed to compare the diagnosticity of various input modalities; investigate whether combining modalities results in superadditive, additive, or redundant effects; clarify the time course of phenomena; model nonlinearity and interactivity among inputs; contrast model predictions with human judgments; and investigate generalizability across people, domains, and contexts. Thus, MLCMs can complement explanation-based approaches, especially for complex, ill-defined phenomena, and they are valuable tools in the arsenal of a pluralistic scientist.

It should also be noted that distinctions among the four main modeling approaches summarized in Figure 2 are not crisp. For example, when theoretical commitments are important, it is possible to prespecify some of the structure and parameters on the basis of theory and/or plausibility while allowing others to be learned (e.g., Hinaut & Dominey, 2013). Similarly, some deep-learning architectures (e.g., convolutional neural networks, which have revolutionized image processing) are inspired by the neural pathways in the visual cortex (Le Cun et al., 2015). When data are abundant but annotations are sparse, a useful approach is to begin with deep representational learning (so that the program automatically learns the features in an unsupervised fashion, i.e., without annotations), but then use standard supervised learning (i.e., with annotations). MLCM development should not be dogmatic; the goals of the enterprise, availability of data, and expertise of the researchers involved should determine the approach.

Illustrative Example

We illustrate the development of an MLCM using Jensen et al.'s (2021) study of teachers' classroom discourse. In this study, audio recordings of teachers' speech were automatically analyzed to estimate the prevalence of seven discourse categories (e.g., questions, elaborated evaluations) linked to students' change in achievement from one grade to another. The main steps to construct the MLCM (which are common to multiple MLCMs) are shown in Figure 3. First, the researchers recorded teachers' audio from 127 authentic class sessions of 16 English Language Arts (ELA) teachers. Next, the recordings were segmented and transcribed into 35,000 utterances via an automatic speech recognizer. Trained coders then annotated 16,000 of these utterances for the presence of each discourse category.

The researchers contrasted two modeling approaches (Fig. 4). The standard approach used utterance-level counts of individual words and two- and three-word phrases (called *n*-grams) as features. Then, binary random forest classifiers (a supervised-learning method) were individually trained to identify the presence/absence of each discourse category on the basis of the features. An examination of the *n*-grams most predictive of each discourse category provided an intuitive understanding of the teachers' talk. For the second approach, the researchers started with a deep neural network that was pretrained on large text corpora containing more than 3 billion words to learn the contextual semantics of words (e.g., to distinguish between "bank" in the context of a river vs. a financial institution) and then fine-tuned the network to identify each discourse category using the 16,000 annotated utterances.

In both approaches, models were evaluated using cross-validation. The utterances were divided into eight partitions; MCLMs were trained on seven partitions (training set) and evaluated for their performance on the held-out partition (test set). The process was repeated until all partitions were included as the test set exactly once. To ensure generalization across teachers, the researchers included utterances of a given teacher in only a training or a testing partition in a given iteration. Accuracy of the models' estimates, defined as correspondence to the human annotations, was somewhat higher for the deep-learning models than for the standard models, which both outperformed chance guessing. The researchers are in the process of embedding the models into a smartphone application that provides teachers with automated feedback on their own classroom discourse to enable reflection and improvement.

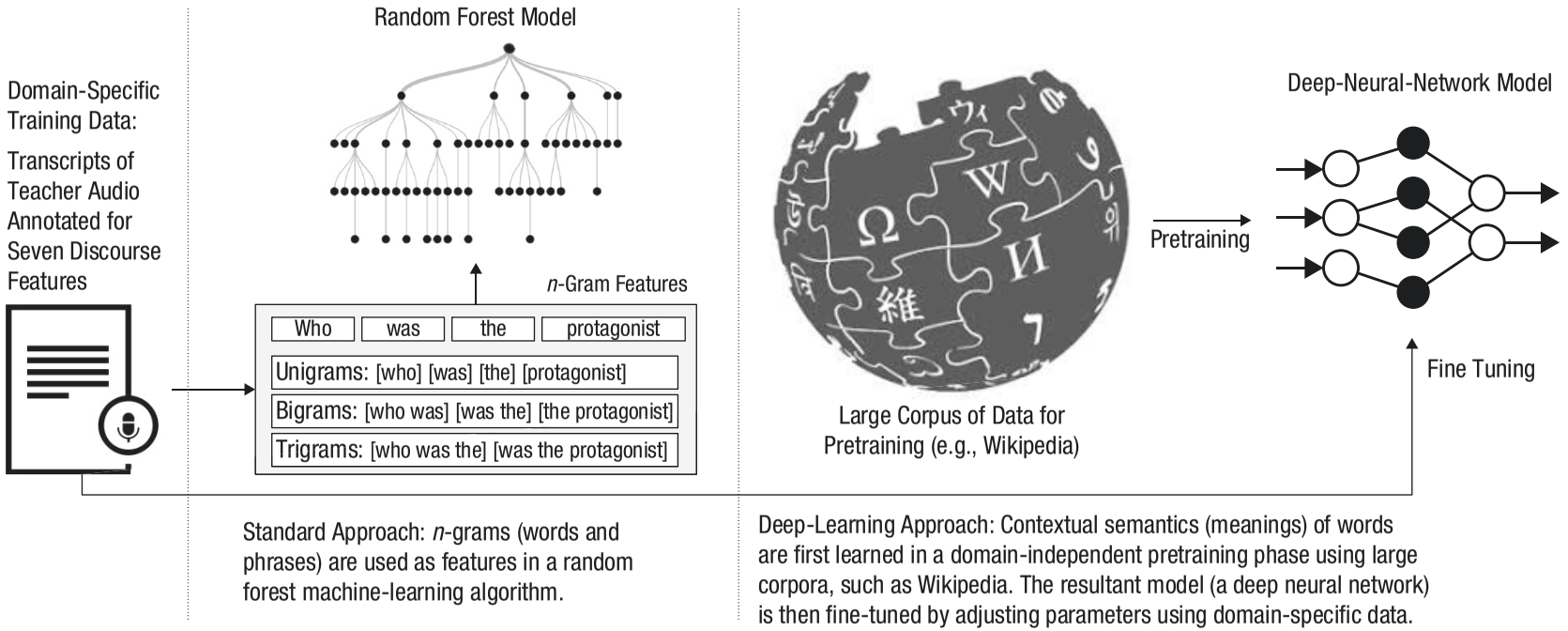

Fig. 4. Standard versus deep-learning approaches for training a machine-learned computational model to classify different types of spoken discourse from audio. In the standard approach, *n*-grams derived from training data are used to produce a random forest model. In the deep-learning approach, contextual semantics are learned from large corpora in a pretraining phase, and the deep neural network is then fine-tuned for the training data.

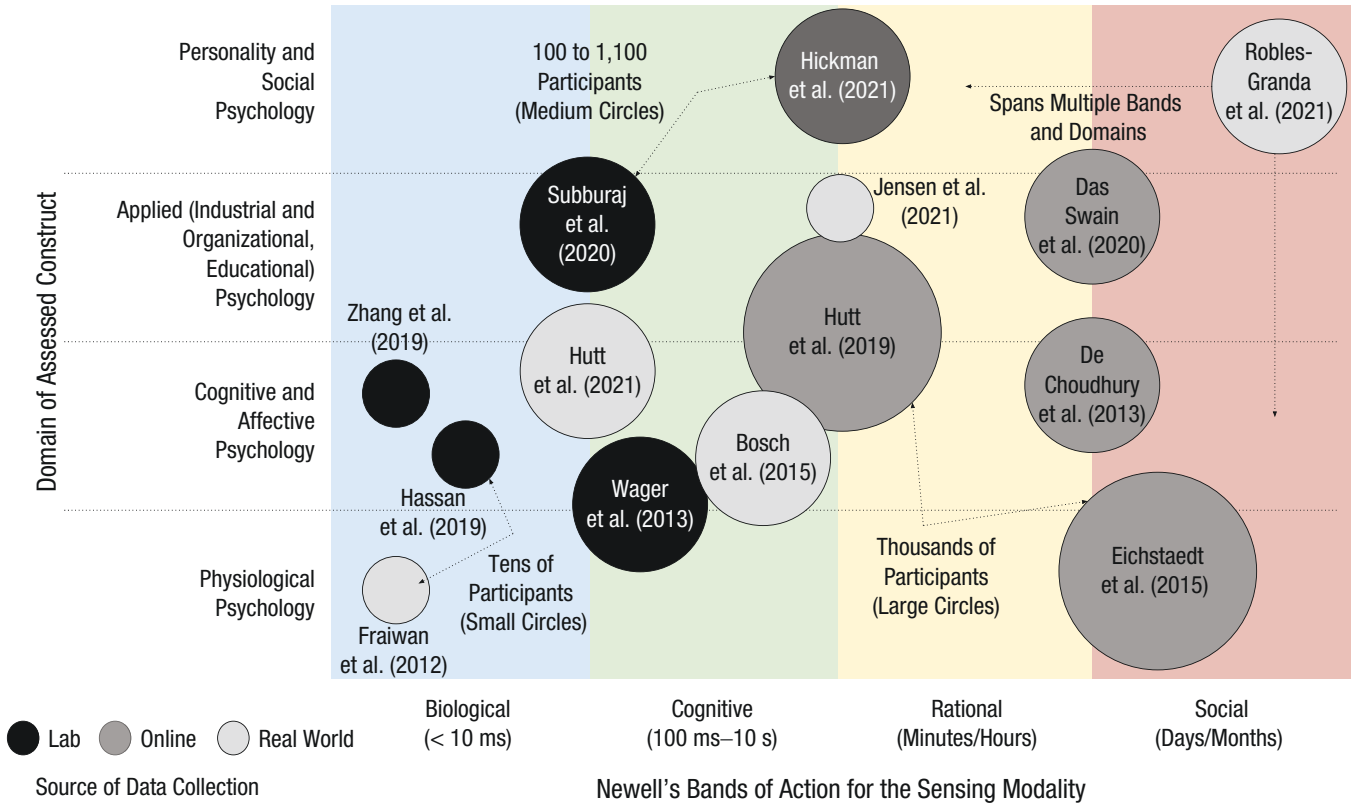

Fig. 5. Selected example cases of machine-learned computational models in four domains of psychological assessment, aligned with respect to Newell's (1990) four bands of action for the input modality and psychological construct assessed. See Table 1 for additional details about the examples.

This example highlights some important points. First, developing MLCMs for complex phenomena, such as classification of spoken discourse, often entails leveraging MLCMs developed for more primitive tasks (e.g., speech recognition, representing word semantics). Second, feature engineering involved minimal human knowledge in that features were automatically computed (standard approach) or bypassed altogether (deep-learning approach). An alternate approach would have been to use handcrafted features, such as parts of speech (e.g., nouns, pronouns), that may have theoretical significance. Third, the results revealed an accuracy-interpretability trade-off: The deep-learning approach yielded more accurate models, and the standard approach yielded more interpretable models.

Selected Examples of MLCMs From the Psychological Sciences

We now present further examples of MLCMs for measurement, which we have roughly organized across four levels of a sensing timescale (i.e., a timescale based on the unit of the input data) inspired by Newell's (1990) bands of action (biological, cognitive, rational, and social; see Fig. 5; Table 1 provides additional details about all the examples). We start with the biological band (sensing interval < 10 ms), which includes some measures of neuronal activity. Fraiwan et al. (2012) developed an MLCM, based on electroencephalography (EEG) data, to accurately discriminate among the five main sleep stages (a time-consuming task for trained clinicians) in a thoracic clinic. In this study, the EEG features were predefined. In contrast, Zhang et al. (2019) used an end-to-end deep-learning approach to develop an MLCM that learned spatiotemporal patterns directly from EEG data to distinguish between high and low workloads. Integrating multiple modalities, Hassan et al. (2019) combined electrodermal activity, photoplethysmography, and electromyography to discriminate among experimentally elicited emotions in the lab.

In a study focused on the cognitive band (100 ms–10 s), Wager et al. (2013) developed an MLCM that discriminates heat-induced pain from warmth, anticipation, recall of pain, and social pain on the basis of whole-brain functional MRI activity.

Whereas the MLCMs in these examples used research-grade sensing and experimentally induced responses in controlled settings, MLCMs can measure spontaneous

Table 1. Additional Details on the Example Case Studies

Study	Sensing band	Signals	Context	*N*	Construct	Level	Machine-learning model
Fraiwan et al. (2012)	Biological	EEG	Clinic	16	Sleep stage	Within individual	Random forest classifier
Zhang et al. (2019)	Biological	EEG	Lab	20	Mental workload	Within individual	Deep neural network
Hassan et al. (2019)	Biological	Physiology (EDA, PPG, EMG)	Lab	32	Affect	Within individual	Deep belief network, support vector machine
Hutt et al. (2021)	Biological/ cognitive	Eye gaze	Classroom	287	Mind wandering	Within individual	Bayesian network
Subburaj et al. (2020)	Biological/ cognitive	Task context, eye gaze, text (speech), facial expressions	Lab	303	Team performance	Within group	Random forest classifier
Wager et al. (2013)	Cognitive	Functional MRI	Scanner	114	Pain	Within individual	Regularized regression
Bosch et al. (2015)	Cognitive	Clicks, task context, facial expressions	Classroom	133	Affect	Within individual	Standard machine learning (various)
Jensen et al. (2021)	Cognitive/ rational	Text (speech)	Classroom	16	Discourse	Within and between individuals	Transformer (deep neural network)
Hickman et al. (2021)	Cognitive/ rational	Text (speech), acoustics, facial expressions	Mostly online, some lab	1,082	Personality	Between individuals	Regularized regression
Hutt et al. (2019)	Cognitive/ rational	Actions (clicks)	Online	69,174	Engagement	Within individuals	Standard machine learning (various)
De Choudhury et al. (2013)	Rational/ social	Text (Twitter)	Online	476	Depression	Between individuals	Support vector machine
Eichstaedt et al. (2015)	Rational/ social	Text (Twitter)	Online	1,347 (counties)	Mortality from heart disease	Between groups	Regularized regression
Das Swain et al. (2020)	Rational/ social	Text (Glassdoor reviews)	Online	341	Organizational culture	Between individuals	Linear regression
Robles-Granda et al. (2021)	Multiple	Text (Facebook), physiology, activity, location, communications metadata without content (e.g., number of phone calls)	Home, office	757	Personality, affect, health, cognitive ability, job performance	Between individuals	Standard machine learning (various)

Note: EEG = electroencephalography; EDA = electrodermal activity; PPG = photoplethysmography; EMG = electromyography.

responses with cost-effective sensing in the real world. Many such studies blend the biological and cognitive sensing bands. For example, Hutt et al. (2021) used $100 eye trackers to develop an MLCM of mind wandering, using data from high-school students while they interacted with educational technology in classrooms. The researchers used the MLCM's estimates to trigger dynamic interventions to reengage attention and improve learning. Similarly, Bosch et al. (2015) combined facial expressions from video with interaction patterns (clicks and click timings) to measure students' affect as they played an educational video game, finding that a multimodal approach improved the model's robustness to missing data but negligibly affected accuracy. Subburaj et al. (2020) used a multimodal (facial expressions, acoustics from speech, eye gaze, and interaction patterns) and multiparty (signals from three individuals) approach to predict collaborative problem-solving outcomes in remote teams.

The rational band consists of measurement in the range of minutes to hours, and studies of activity in this band often aggregate more fine-grained sensing (cognitive band) over longer time frames (rational band). Jensen et al.'s (2021) study of teachers' discourse, discussed above, is one example. Another is Hickman et al.'s (2021) study, in which language, facial expressions, and prosody in mock video interviews for personnel selection were used to develop an automated system for scoring personality. In a large-scale study, Hutt et al. (2019) developed an MLCM to infer engagement from interaction patterns of approximately 70,000 students as they interacted with an online learning platform.

Studies at the social band have largely relied on social-media posts using time frames from days to months (individual posts are in the rational band, so this work entails combining the rational and social bands). De Choudhury et al. (2013) developed an MLCM that identified individuals diagnosed with depression on the basis of their Twitter usage. Eichstaedt et al. (2015) also used Twitter data, but at the societal level, to predict county-level rates of mortality from atherosclerotic heart disease. Their MLCM was a better predictor than established demographic and health indicators (but see Brown & Coyne, 2018, for an alternate interpretation). At the organization level, Das Swain et al. (2020) analyzed language used in more than 600,000 Glassdoor reviews of 92 Fortune 500 companies to infer 41 dimensions of organizational culture, which then were used to predict job performance.

MLCMs can span all four bands. In a yearlong study of 757 information workers, Robles-Granda et al. (2021) measured physical and physiological signals from wearable sensors, communications from a smartphone app, relative location based on Bluetooth beacons, contextual cues (e.g., weather), and social-media data to develop MLCMs of personality, cognitive ability, affect, health, and job performance.

Accuracy and Generalizability of MLCMs

MLCMs are typically evaluated for their accuracy and generalizability. Accuracy is higher with well-engineered features and sophisticated algorithms that can infer complex patterns without overfitting; simpler approaches risk underfitting the data. It is often assumed that big data yield higher accuracies than smaller data sets, but this is an oversimplification; it is not the volume that matters but rather the quality of the data and how well they represent the phenomenon to be measured. Another assumption is that multimodality improves accuracy, but this is not always the case (e.g., the Bosch et al., 2015, study); often its main advantage is increasing robustness (D'Mello & Kory, 2015). All things being equal, the quality of the annotations matters most because it provides the supervisory signal for learning and evaluation. High-quality annotations should reach the same standards of construct validation used for any psychological measure (e.g., reliability, convergent validity).

MLCMs developed in very specific contexts are unlikely to generalize beyond the specific paradigm (e.g., Hassan et al.'s, 2019, affect-induction study), though this lack of generalizability can be somewhat alleviated by training on multiple stimuli or tasks (e.g., Zhang et al., 2019, used both spatial and arithmetic tasks). Hutt et al. (2019) made domain generalizability a design principle in selecting features for their engagement study, and their MLCM trained on algebra data generalized to geometry data without retraining. Temporal generalizability is of concern for language models as new terms enter the lexicon. The gold standard is to collect broad, diverse, and voluminous training data, as was done in the example studies using data from social media, but this is challenging for sensor-based models (e.g., Robles-Granda et al.'s, 2021, study) without mass surveillance. Starting with models pretrained on large data sets across multiple domains and customizing them using limited data in a target domain (as in the Jensen et al., 2021, example) is a promising approach.

Expectations of accuracy and generalizability must be calibrated with respect to the complexity of the construct, the underlying context, and the availability of good-quality training data (especially annotations). Accuracy will be higher for well-defined, experimentally induced phenomena in the context of the lab (Wager et al.'s, 2013, pain example) than for spontaneously occurring, ill-defined phenomena across multiple contexts (Hutt et al., 2019, used online data collected

in classrooms, homes, etc., in their engagement study). Similarly, generalizability is difficult when the phenomenon is highly context-specific (e.g., emotion; D'Mello et al., 2018). In such cases, it is prudent to use context-specific models, live with modest accuracy, and temper performance claims rather than completely write off the approach. We suggest channeling Tukey (1962) when interpreting the value of such models: "far better an approximate answer to the right question, which is often vague, than the exact answer to the wrong question, which can always be made precise" (p. 13).

Bias, Fairness, and Interpretability of MLCMs

There was a media frenzy a few years ago when it was revealed that commercial face-recognition technology routinely underperforms for dark- compared with lighter-skinned individuals—with shocking disparities (error rates up to 34.7% vs. 0.8%; Buolamwini & Gebru, 2018). Although the idea of biased algorithms dates back to the 1970s, similar high-profile revelations have recently renewed interest in raising awareness of algorithmic bias and approaches to mitigate it.

We (Tay et al., 2021) have proposed a theoretical framework for addressing bias in MLCMs used for psychological assessment. In this framework, systematic departures of some subgroups' MLCM scores from their actual scores is evidence for *bias* if there are no actual subgroup differences. Though the terms are often used interchangeably, *fairness* is distinct from bias. It is a subjective perspective based on the values and beliefs of individuals and societies.

For example, consider an MLCM that assesses personality using automated video interviews, which are increasingly used in real-world hiring (e.g., Hickman et al., 2021). If the MLCM yields higher conscientiousness scores for men than for women and nonbinary individuals but there are no gender differences in the annotations used to train the model (e.g., expert-rated conscientiousness), this would be prima facie evidence of bias. On the other hand, if the annotations for agreeableness indicate higher scores for men than for women and nonbinary individuals and the MLCM reproduces this pattern (i.e., it is not biased), some people would view the MLCM as fair because its measurements reflect actual scores. Others would view it as unfair because it gives unequal group outcomes.

It is sometimes assumed that bias is purely a reflection of the representativeness of the data used to develop MLCMs, but in fact, it arises from decisions made throughout the modeling process. Our framework identifies and contextualizes potential sources of bias at both the data and the algorithm levels while also recommending tests and mitigation strategies.

A related concept is *interpretability* (or *explainability*), the degree to which the inner workings of the model are interpretable by humans, a critical concern for both scientific inquiry and real-world use. Explainability can pertain to the structure of an MCLM itself (e.g., how do the features combine? what are the representations?) and/or to the MLCM's outputs (e.g., why did the model predict X for data point Y?). The four modeling approaches in Figure 2 align along an interpretability-performance continuum, with the handcrafted models and deep-learning approaches on either extreme. Whereas methods from the nascent field of explainable artificial intelligence (XAI) can help improve the interpretability of MLCMs (e.g., Lundberg et al., 2020), it is unlikely that the trade-off will be entirely eliminated (cf. the no-free-lunch theorem of mathematical folklore).

MLCMs in a Well-Measured Life

What role do MLCMs play in an era obsessed with measurement? As the examples illustrate, MLCM-based measures have been developed in multiple areas of psychological sciences, ranging from neuroscience to cognitive and affective science, education, organizational culture, and personality and social psychology (Fig. 5). They reflect measurements in the scanner, in the lab, online, at workplaces, at homes, in schools, and in the community. Whereas most MLCMs focus on within- and between-individuals differences, some produce measurements at the level of the team, organization, or society. MLCMs have been used for scientific inquiry, automated scoring, assessment, and intervention. They extend researchers' capacity to harness natural data sources, in each case greatly increasing the speed, scale, and convenience of psychological measurement. Psychological scientists have a vital role to play in the future of MLCMs by providing guidance on human behavior, construct validity, statistical rigor, theoretical grounding, and evaluations of bias and fairness.

At the same time, a proliferation of such measures increases privacy, security, and ethical concerns over what and how data are collected, processed, and stored, as well as the purpose for which they are collected and analyzed. It also raises long-established concerns of bias and fairness. Whereas researchers have historically emphasized accuracy and generalizability, the idea of achieving unbiased, fair, and interpretable models has garnered considerable interest over the past decade. As research and recommendations emerge, one immediate step should be to adopt a culture in which ethical design is a core goal. For example, the National Science Foundation National AI Institute on Student-AI Teaming[1] has adopted a Responsible Innovation Framework

(Stilgoe et al., 2013) that guides its vision, values, methods, and criteria for success. Of course, words must be followed by action so that the products of research (including MLCMs) are instruments that reflect and promote justice rather than perpetuate inequality.

Recommended Reading

D'Mello, S., Kappas, A., & Gratch, J. (2018). (See References). Provides a general tutorial on how to construct machine-learned computational models in the domain of emotion.

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology, 14*, 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037. Reviews the use of machine learning for assessment in clinical psychology.

Jordan, M. I., & Mitchell, T. M. (2015). (See References). Provides an accessible tutorial on machine learning and a review of recent advances.

Tay, L., Woo, S. E., Hickman, L., Booth, B., & D'Mello, S. (2021). (See References). Provides a framework integrating psychometric concepts of bias with machine learning for the purpose of psychological assessment.

Yarkoni, T., & Westfall, J. (2017). (See References). Distinguishes between building predictive and explanatory models in psychology.

Transparency

Action Editor: Robert L. Goldstone
Editor: Robert L. Goldstone
Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This research was supported by the National Science Foundation (DRL 2019805, DRL 1920510, SES 2030599, SES 1928612, SES 1921111, IIS 1921087, IIS 1523091/1748739, and DUE 1745442/1660877) and by the Institute of Education Sciences (R305A170432 and R305C160004). Any opinions, findings and conclusions, or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the funding agencies.

ORCID iDs

Sidney K. D'Mello https://orcid.org/0000-0003-0347-2807
Louis Tay https://orcid.org/0000-0002-5522-4728

Note

1. The National AI Institute for Student-AI Teaming (www.isat.ai) is one of the seven inaugural National Artificial Intelligence Institutes.

References

Bosch, N., Chen, H., D'Mello, S., Baker, R., & Shute, V. (2015). Accuracy vs. availability heuristic in multimodal affect detection in the wild. In *ICMI '15: Proceedings of the 2015 ACM International Conference on Multimodal Interaction* (pp. 267–274). Association for Computing Machinery. https://doi.org/10.1145/2818346.2820739

Brown, N. J. L., & Coyne, J. C. (2018). Does Twitter language reliably predict heart disease? A commentary on Eichstaedt et al. (2015a). *PeerJ*, *6*, Article e5656. https://doi.org/10.7717/peerj.5656

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, *81*, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

Card, S. K., Moran, T., & Newell, A. (1983). *The psychology of human-computer interaction*. Erlbaum.

Das Swain, V., Saha, K., Reddy, M. D., Rajvanshy, H., Abowd, G. D., & De Choudhury, M. (2020). Modeling organizational culture with workplace experiences shared on Glassdoor. In *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery. https://doi.org/10.1145/3313831.3376793

De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media* (pp. 128–137). Association for the Advancement of Artificial Intelligence. https://ojs.aaai.org/index.php/ICWSM/article/view/14432/14281

D'Mello, S., Kappas, A., & Gratch, J. (2018). The affective computing approach to affect measurement. *Emotion Review*, *10*(2), 174–183. https://doi.org/10.1177/1754073917696583

D'Mello, S. K., & Kory, J. (2015). A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys*, *47*(3), Article 43. https://doi.org/10.1145/2682899

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., Jha, S., Agrawal, M., Dziurzynski, L. A., Sap, M., Weeg, C., Larson, E. E., Ungar, L. H, & Seligman, M. E. P. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, *26*(2), 159–169. https://doi.org/10.1177/0956797614557867

Faber, M., Bixler, R., & D'Mello, S. K. (2018). An automated behavioral measure of mind wandering during computerized reading. *Behavior Research Methods*, *50*(1), 134–150. https://doi.org/10.3758/s13428-017-0857-y

Fraiwan, L., Lweesy, K., Khasawneh, N., Wenz, H., & Dickhaus, H. (2012). Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier. *Computer Methods and Programs in Biomedicine*, *108*(1), 10–19. https://doi.org/10.1016/j.cmpb.2011.11.005

Hassan, M. M., Alam, M. G. R., Uddin, M. Z., Huda, S., Almogren, A., & Fortino, G. (2019). Human emotion recognition using deep belief network architecture. *Information Fusion*, *51*, 10–18. https://doi.org/10.1016/j.inffus.2018.10.009

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2021). Automated video interview personality assessments: Reliability, validity, and generalizability investigations.

Journal of Applied Psychology. Advance online publication. https://doi.org/10.1037/apl0000695

Hinaut, X., & Dominey, P. F. (2013). Real-time parallel processing of grammatical structure in the fronto-striatal system: A recurrent network simulation study using reservoir computing. *PLOS ONE*, *8*(2), Article e52946. https://doi.org/10.1371/journal.pone.0052946

Hutt, S., Grafsgaard, J. F., & D'Mello, S. K. (2019). Time to scale: Generalizable affect detection for tens of thousands of students across an entire school year. In *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Paper 496). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300726

Hutt, S., Krasich, K., Brockmole, J. R., & D'Mello, S. K. (2021). Breaking out of the lab: Mitigating mind wandering with gaze-based attention-aware technology in classrooms. In *CHI '21: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 52). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445269

Jensen, E., Pugh, S. L., & D'Mello, S. K. (2021). A deep transfer learning approach to modeling teacher discourse in the classroom. In *LAK21: Proceedings of the 11th Learning Analytics and Knowledge Conference* (pp. 302–312). Association for Computing Machinery. https://doi.org/10.1145/3448139.3448168

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Le Cun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Newell, A. (1990). *Unified theories of cognition*. Harvard University Press.

Robles-Granda, P., Lin, S., Wu, X., Martinez, G. J., Mattingly, S. M., Moskal, E., Striegel, A., Chawla, N. V., D'Mello, S., Gregg, J., Nies, K., Mark, G., Grover, T., Campbell, A. T., Mirjafari, S., Saha, K., De Choudhury, M., & Dey, A. K. (2021). Jointly predicting job performance, personality, cognitive ability, affect, and well-being. *IEEE Computational Intelligence Magazine*, *16*(2), 46–61. https://doi.org/10.1109/MCI.2021.3061877

Stilgoe, J., Owen, R., & Macnaghten, P. (2013). Developing a framework for responsible innovation. *Research Policy*, *42*(9), 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

Subburaj, S. K., Stewart, A. E. B., Rao, A. R., & D'Mello, S. K. (2020). Multimodal, multiparty modeling of collaborative problem solving performance. In *ICMI '20: Proceedings of the 2020 International Conference on Multimodal Interaction* (pp. 423–432). Association for Computing Machinery. https://doi.org/10.1145/3382507.3418877

Tay, L., Woo, S. E., Hickman, L., Booth, B., & D'Mello, S. (2021). *A conceptual framework for investigating and mitigating machine learning bias for psychological assessment* [Manuscript submitted for publication]. Department of Psychological Sciences, Purdue University.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*(1), 1–67. https://doi.org/10.1214/aoms/1177704711

Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C.-W., & Kross, E. (2013). An fMRI-based neurologic signature of physical pain. *New England Journal of Medicine*, *368*(15), 1388–1397. https://doi.org/10.1056/NEJMoa1204471

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Routledge.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zhang, P., Wang, X., Zhang, W., & Chen, J. (2019). Learning spatial–spectral–temporal EEG features with recurrent 3D convolutional neural networks for cross-task mental workload assessment. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *27*(1), 31–42. https://doi.org/10.1109/TNSRE.2018.2884641